肽段酶切概率预测软件使用说明书

Version 1.7.0

Last revised February 13, 2023

杨婧涵[1,3‖]，高志强[1,3‖]，任修涵[4]，盛捷[2]，徐平[2]，常乘[2*]，付岩[1,3*]

[1]CEMS, NCMIS, RCSDS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.

[2]State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, Beijing 102206, China.

[3]School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China.

[4]School of Sciences, China University of Mining & Technology, Beijing 100083, China.

‖共同第一作者

*通讯作者:

付岩，邮箱：yfu@amss.ac.cn

常乘，邮箱：changchengbio@163.com

1 软件介绍

如今，基于高通量生物质谱技术的蛋白质组学已经成为生物学、医学领域研究的一种前沿方法。在主流的鸟枪法蛋白质组学分析流程中，蛋白质水解产生的肽段将经由质谱仪进行检测，再通过对肽段质谱数据的分析完成对蛋白质的定性和定量。[1] 通常，不同的水解酶可能会在相应的特异性氨基酸位点处发生酶切，但是漏切的现象普遍存在，从而影响了部分肽段的生成，导致其不能被质谱检测到，最终会阻碍了人们对质谱数据进行高精度、大规模地解析。过去，已有一些关于预测肽段酶切概率的研究，但都基于经验规则或传统的机器学习模型，预测精度不尽如人意，并且仅针对少数几种酶，主要是胰蛋白酶（trypsin），难以满足实际应用的需求。因此，若能准确预测多种常用酶的特异性位点发生酶切的概率，将有助于改善蛋白质组学的实验设计和数据分析。于是，我们开发了一种基于深度学习的支持八种常用水解酶的肽段酶切概率预测软件 DeepDigest。与传统的机器学习算法以及已发表的肽段酶切概率预测软件 MC:pred[2] 相比，该软件具有更强的学习和泛化能力。仅需输入蛋白质序列，该软件就可以自动完成对蛋白质序列的理论酶切和对肽段酶切概率的准确预测。DeepDigest 是一个命令行工具，可以从如下网址免费下载使用：http://fugroup.amss.ac.cn/software/DeepDigest/DeepDigest.html。

2 软件使用环境

2.1 硬件环境

- CPU：2.2GHZ 以上
- 内存：2G 以上
- 硬盘：100G 以上

2.2 软件环境

- 已经验证的操作系统
  - Windows 7（64 位）

- Windows 10（64 位）

- Python 3.5 版本
  - TensorFlow 1.10.0 版本库
  - Keras 2.2.4 版本库
  - Numpy 1.14.5 版本库

3 软件使用手册

3.1 软件分析流程

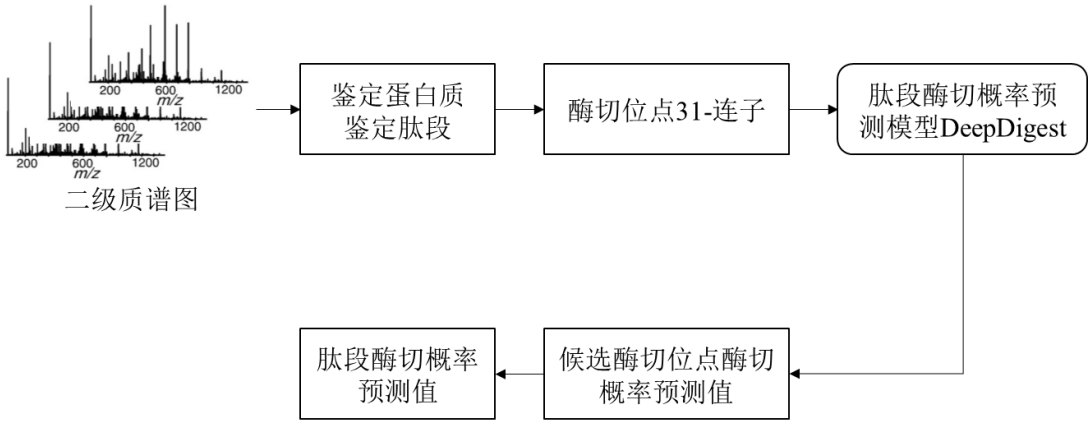

图 1 肽段酶切概率预测软件流程图

软件分析流程如图 1 所示。首先，通过 MaxQuant[3,4] 或其他搜库鉴定软件 [5–9] 对二级质谱图进行蛋白质和肽段的鉴定分析。然后，根据指定水解酶的特异性对鉴定蛋白质的所有候选酶切位点提取 31-连子（包含位点），即分别取位点 N 端和 C 端的 15 位氨基酸加上位点本身所构成的序列，构建酶切位点 31-连子的数据集。接着，将数字化编码后的所有 31-连子输入到肽段酶切概率预测模型 DeepDigest 中，由词嵌入层习得氨基酸的分布式表示，再由卷积和平均池化操作提取序列局部特征，并利用长短时记忆网络捕捉序列长时依赖特征，最终输出对候选酶切位点酶切概率的预测值，并计算所有肽段酶切概率的预测值。

3.2 软件环境配置

DeepDigest 是一个基于 Python 语言的命令行工具，在使用时需要满足对应

版本的 Python 环境配置。为了方便用户使用，建议直接在如下官方网址下载并安装 Anaconda：https://www.anaconda.com/。下载页面如图 2 所示：

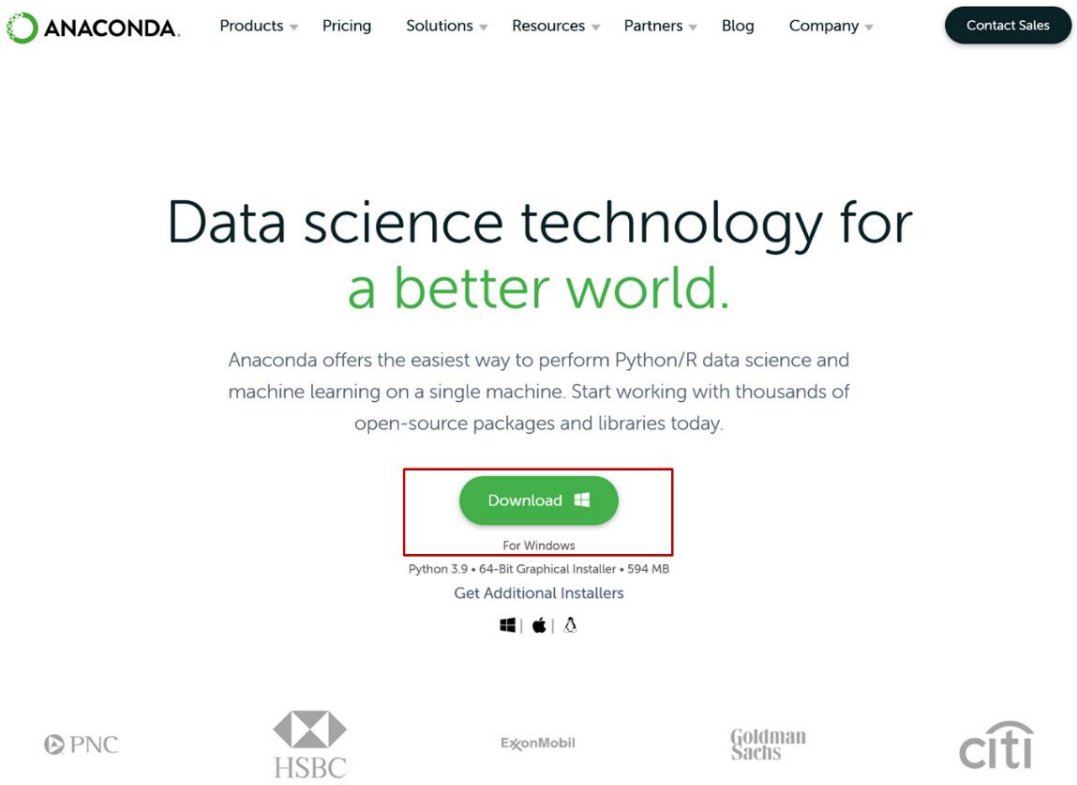

图 2 官网下载 Anaconda 示意图

双击用户使用的电脑操作系统对应的下载文件即可自动安装，并自动配置系统环境变量。但是，还需注意的是，官网上下载的 Anaconda 通常只会自动安装最新版本的 Python 环境，用户还需要自行创建运行 DeepDigest 所对应的 Python 版本。安装 Anaconda 的一大便利之处就在于，只需简单几个步骤就可以完成多版本 Python 环境的配置，可供用户自由切换。

首先，点击桌面左下角“开始”按钮，键入“cmd.exe”打开命令行解释器。然后，输入“conda --version”命令查看当前的 conda 版本，并确认当前安装的 Anaconda 为最新版（图 3a）。若不确定是否为最新版本，可通过“conda update conda”和“conda update anaconda”两个命令进行升级（图 3b）。

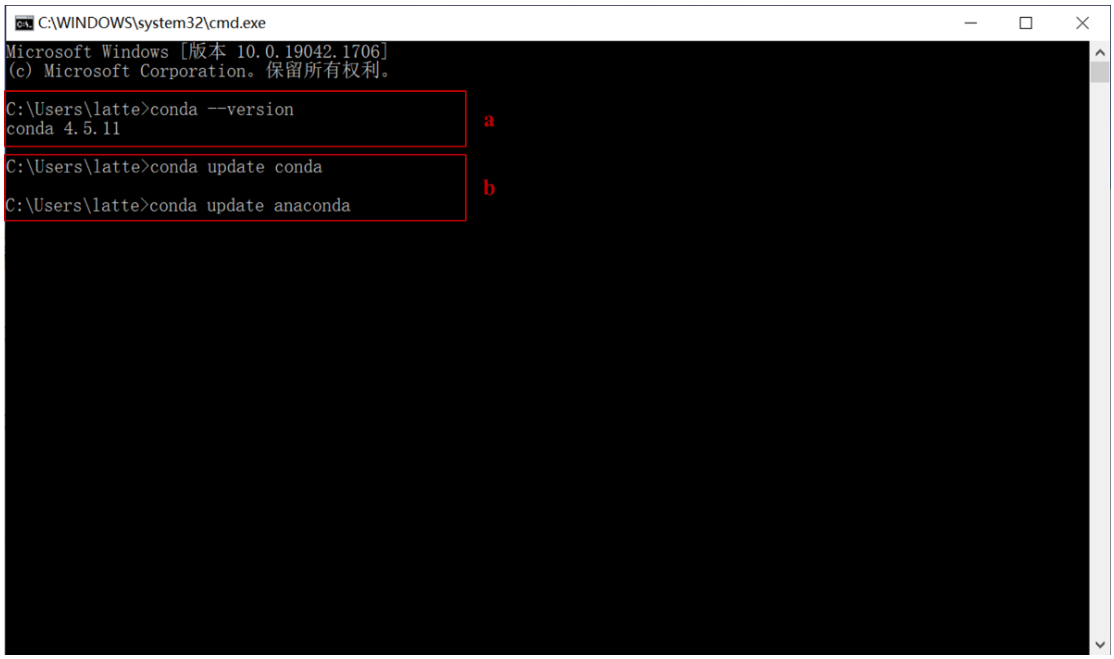

图 3 查看并升级 Anaconda 版本示意图

接着，使用“conda create --name python3.5 python=3.5”命令创建一个新的 Python 3.5 版本的环境（图 4）。

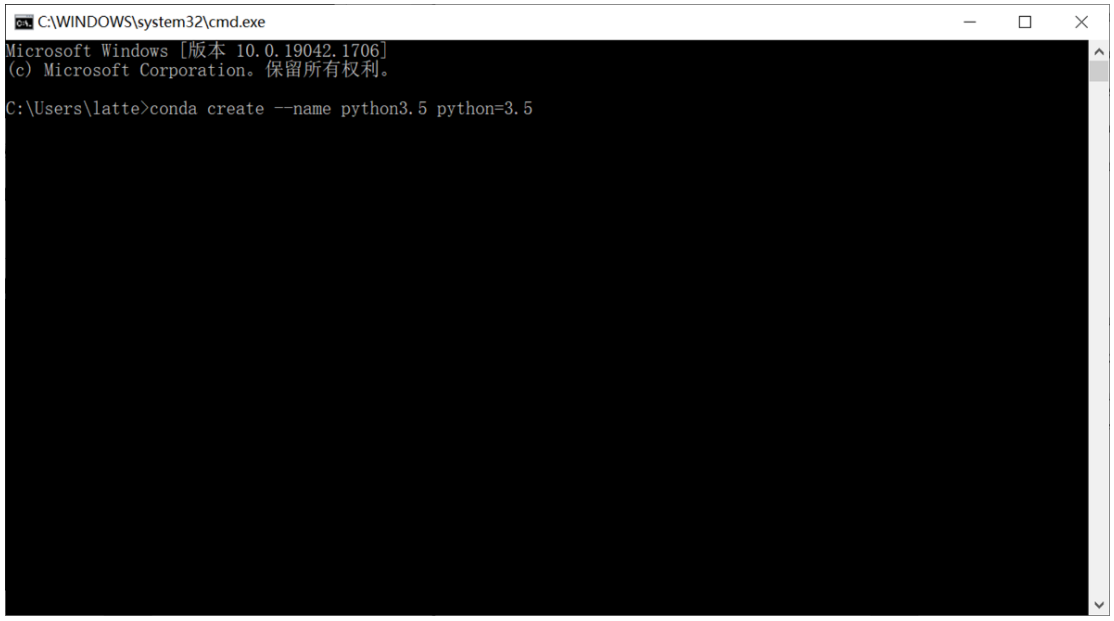

图 4 创建新版本 Python 环境示意图

创建完成后，可以通过“conda info -e”命令查看当前已经安装的所有 Python 环境（图 5a）。在使用 DeepDigest 前，只需要在命令行解释器中使用“activate python3.5”命令激活 Python 3.5 环境即可（图 5b）。

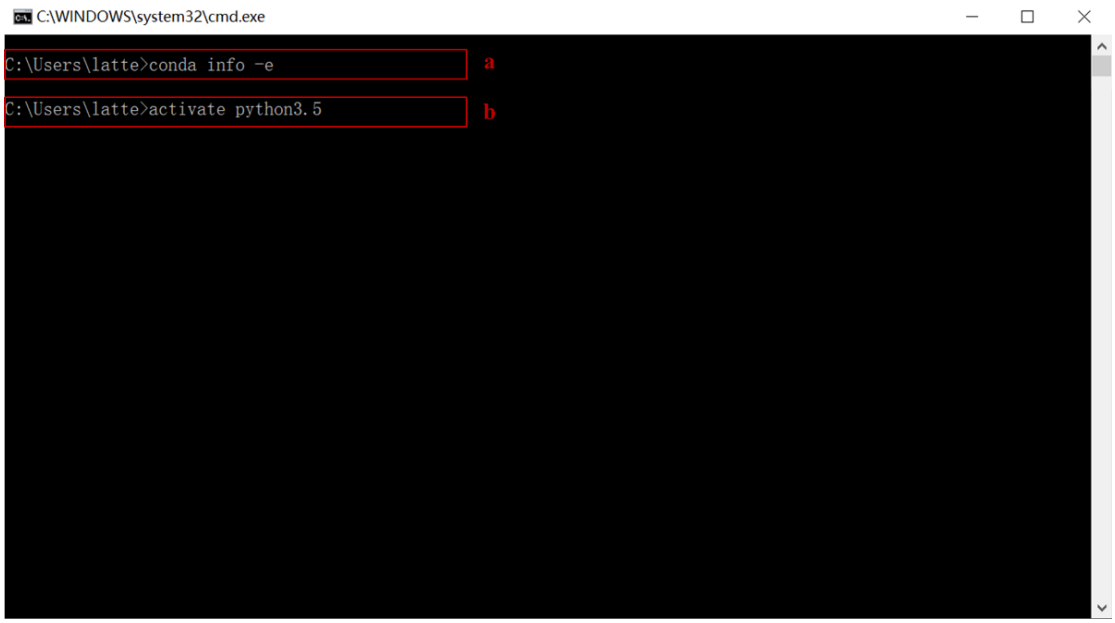

图 5 查看所有 Python 环境并激活新版本 Python 环境示意图

3.3 使用命令行运行软件

3.3.1 参数设置

使用 DeepDigest 时，用户可以自定义一系列参数设置，具体参数及其含义如下表所示：

表 1 肽段酶切概率预测软件 DeepDigest 的参数说明

参数名称	含义
input	蛋白质序列文件路径（.fasta）
output	输出文件路径（.txt）
regular	在蛋白质序列文件中提取蛋白质标识名的正则表达式（默认值：">(.*?)\s"）
protease	水解酶类型（默认值：Trypsin）
missed_cleavages	蛋白质理论酶切肽段的允许最大漏切位点数（默认值：2）
min_len	蛋白质理论酶切肽段的允许最小长度（默认值：7）

max_len	蛋白质理论酶切肽段的允许最大长度（默认值：47）

3.3.2 运行命令

首先，点击桌面左下角“开始”按钮，键入“cmd.exe”回车打开命令行解释器。然后，通过 cd 命令设置 the_main.py 所在目录为当前工作目录。最后，按照如下格式调用 DeepDigest：

python the_main.py --input=*the path of protein sequence file* --output=*the path of output file* --regular=">(.*?)\s" --protease=Trypsin --missed_cleavages=2 --min_len=7 --max_len=47

图 6 为一个实例：

```
E:\DeepDigest>python the_main.py --input=E:\DeepDigest\nextprot-sparql-entry_PE3.fasta
 --output=E:\DeepDigest\PredictResultsOfPE3_Trypsin.txt --regular="">(.*?)\s" --proteas
e=Trypsin --missed_cleavages=2 --min_len =7 --max_len=47
```

图 6 命令行调用 DeepDigest 实例

3.4 结果文件说明

DeepDigest 运行结束后会在用户指定的目录下生成一个结果文件（.txt），表 2 为结果文件内容说明。

表 2 DeepDigest 结果文件说明

列名	含义
Protein id	蛋白质标识名
Peptide sequence	蛋白质的理论酶切肽段
Digestibility of N-terminal site	理论酶切肽段 N 端酶切位点的酶切概率预测值
Digestibility of C-terminal site	理论酶切肽段 C 端酶切位点的酶切概率预测值
Digestibility of the missed site(s)	理论酶切肽段所有漏切位点的酶切概率预测值

4 参考文献

1. Altelaar, A. F. M., Munoz, J. & Heck, A. J. R. Next-generation proteomics: Towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* **14**, 35–48 (2013).

2. Lawless, C. & Hubbard, S. J. Prediction of missed proteolytic cleavages for the selection of surrogate peptides for quantitative proteomics. *Omi. A J. Integr. Biol.* **16**, 449–456 (2012).

3. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).

4. Cox, J. *et al.* Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).

5. Park, C. Y., Klammer, A. A., Käli, L., MacCoss, M. J. & Noble, W. S. Rapid and accurate peptide identification from tandem mass spectra. *J. Proteome Res.* **7**, 3022–3027 (2008).

6. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: An open-source MS/MS sequence database search tool. *Proteomics* **13**, 22–24 (2013).

7. Chi, H. *et al.* PFind-Alioth: A novel unrestricted database search algorithm to improve the interpretation of high-resolution MS/MS data. *J. Proteomics* **125**, 89–97 (2015).

8. Chi, H. *et al.* Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nat. Biotechnol.* **36**, 1059–1066 (2018).

9. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14**, 513–520 (2017).